Project Proposal

Harshitha Manduva(u1080236)

Paridhi Bahethi(u1142877)

Paper title and source:

De-anonymizing networks: http://www.ra.ethz.ch/CDStore/www2007/www2007.org/papers/paper232.pdf

Algorithm and the problem being solved:

Many social networking websites make their data public after anonymizing it. The paper talks about an algorithm that can violate the privacy of users from this anonymized social network. The privacy can be violated in both active and passive approach. According to the paper in an active approach we create fake user accounts to get the data. But in the passive approach, we observe the data with the immediate circle of the attacker. We are planning to implement the walk base active attack. The passive approach is similar but we will not be creating a subgraph with fake accounts but rather search with the existing subgraph knowledge we have. The active attack goes as follows:

Setup required before releasing the anonymous graph:

1. Create k fake users and connect them to other fake users using Hamiltonian paths.
2. Decide on the number of accounts that are to be compromised (target accounts).
3. Connect the fake accounts with the target accounts.
4. We connect each target unique fake accounts. In order to make these accounts trackable.

Recovery of nodes from the graph:

1. Search through the graph to locate our own fake accounts. This is done by looking for degrees and Hamiltonian paths.
2. Once we get the subgraph, we can identify the targets in time proportional to the size of this subgraph.
3. Now that we have identified the targets, their data can be easily compromised.

Datasets planning to use:

The primary criteria for choosing a data set for us is the undirected social network. And the below 2 are the freely available undirected masked social network data i.e., data of facebook and data of Enron email by stanford. Below are the links, we are planning to take any one of these 2 data sets:
https://snap.stanford.edu/data/egonets-Facebook.html
https://snap.stanford.edu/data/email-Enron.html

Plans on how to improve the algorithm:

Improving the masked graph to improve privacy:

Changing the masked graph that will be released in such a way that this type of attacks can be prevented, but the overall structure can still be preserved. By modifying the graph structure, we will compromise a bit on the integrity of the structure to improve the privacy. And this trade off exists. So For improving the privacy of the masked graph, we can use strategies like: add random nodes to the original graph, remove some of the random edges in the graph or introduce random edges inside the graph. This may significantly decrease the success rate of this type of attacks. How many nodes and edges must be manipulated to still preserve the structure of the graph can be determined with some trial and error and plotting the results.

Improving detection of subgraph in the masked graph:

On an initial thought, the algorithms performance can be improved by changing the hyper parameters in the algorithm. Like the number of fake accounts to be created (k). And reducing the values of the parameters [d0, d1] which determines the fake nodes' degrees. And instead of using a constant value C to connect the target node to the number of fake accounts, the value of C can also be made to choose random from the set of hyperparameters say [c0, c1]. This makes the subgraph that we are trying to find in the masked data more unique and easier to identify.